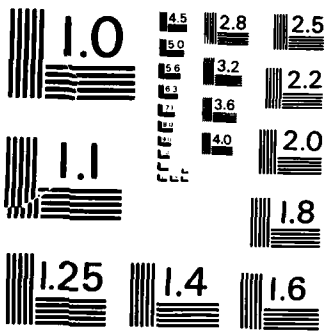
AD-A163 601 BIAS AND EFFICIENCY OF THE CONSISTENT WEIGHTED REGRESSION ESTIMATORS IN F (U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER L DENG SEP 85 MRC-TSR-2870 1/1

UNCLASSIFIED DAAG29-80-C-0041 F/G 12/1 NL



(2)

MRC Technical Summary Report #2870

BIAS AND EFFICIENCY OF THE CONSISTENT WEIGHTED REGRESSION ESTIMATORS IN FINITE POPULATION SAMPLING

Lih-Yuan Deng

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

September 1985

(Received August 26, 1985)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

BIAS AND EFFICIENCY OF THE CONSISTENT WEIGHTED REGRESSION ESTIMATORS IN FINITE POPULATION SAMPLING

Lih-Yuan Deng*

Technical Summary Report #2870

September 1985

ABSTRACT

The leading terms of the bias of the ratio and regression estimators are known to be of order n^{-1}. We use a finite population decomposition to give a different expression for the leading term of the bias. Fitting a regression line to the finite population, we show that the intercept of the regression line causes the bias of the ratio estimator. Fitting a quadratic regression to the finite population, we show that the bias of the regression estimator is caused by the quadratic term. We also give a compact and intuitive formula for the leading term of the bias of the weighted regression estimators for p-auxiliary variables. Using the same decomposition, we can rewrite the variance formula of some popular estimators in terms of some simple and interpretable population characteristics. We prove that under simple random sampling scheme the unweighted regression estimator is the most efficient estimator. The extension for the p-auxiliary variates is also given.

AMS (MOS) Subject Classification: 62D05

Key Words: Ratio estimator; Regression estimator; Weighted regression estimator; Bias; Efficiency; Finite population decomposition.

Work Unit Number 4 - Statistics and Probability

* Assistant Professor, Department of Mathematical Sciences, Memphis State University, Memphis, TN. 38152.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

In survey sampling, we often make use of the auxiliary covariate to improve the precision of estimating the population mean of a character of interest. Ratio and regression estimators are two commonly used estimators. It is well-known that they have a small order bias. We give a new interpretation of the bias using a finite population decomposition. Fitting a regression line to the finite population, we show that the intercept of the regression line causes the bias of the ratio estimator. Fitting a quadratic regression to the finite population, we show that the bias of the regression estimator is caused by the quadratic term. We also give a compact and intuitive formula for the leading term of the bias of the weighted regression estimators for p-auxiliary variables. We also prove that under simple random sampling scheme the unweighted regression estimator is the most efficient estimator.

BIAS AND EFFICIENCY OF THE CONSISTENT WEIGHTED REGRESSION ESTIMATORS IN FINITE POPULATION SAMPLING

Lih-Yuan Deng*

1. Introduction

Consider a finite population consisting of N units with values (y_i, x_i), i=1,2,...,N, where x_i is positive and known. A simple random sample of size n is chosen without replacement from the population. Denote the sample and population means of y and x by $\bar{y}, \bar{x}$ and $\bar{Y}, \bar{X}$ respectively. The ratio estimator $\hat{\bar{y}}_R = \bar{X}\,\bar{y}/\bar{x}$ and the linear regression estimator $\hat{\bar{y}}_{lr} = \bar{y} - b(\bar{x} - \bar{X})$ are the most commonly used estimators of $\bar{Y}$, where $b = \sum_1^n(x_i - \bar{x})(y_i - \bar{y})/\sum_1^n(x_i - \bar{x})^2$ is the sample regression coefficient of y over x.

The ratio estimator is known to have a bias of order n^{-1} . Durbin(1959), Beale(1962) and Tin(1965) proposed several estimators to reduce its n^{-1} bias. In Section 2, we use a finite population decomposition to give a different expression of the leading term of the bias. Fitting a regression line of y over x to the finite population, we show that the intercept of the regression line causes the leading term of the bias.

Like the ratio estimator, the linear regression estimator is also a biased estimator. To study the bias of the regression estimator we introduce a different decomposition. By fitting a quadratic regression to the finite population, we show that the leading term of the bias is caused by the quadratic term. The sign of the bias is also

* Assistant Professor, Department of Mathematical Sciences, Memphis State University, Memphis, TN. 38152.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

determined by the coefficient of the quadratic term. In fact, we show that a negative (positive) coefficient indicates a slight overestimation (underestimation) of $\hat{\bar{y}}_{lr}$.

Konijn (1973) gave an explicit but complicated expression of the leading term of the bias for bivariate regression estimator. In Section 4, we give a compact and more intuitive formula for the leading term of the bias of the p dimensional weighted regression estimators.

The underlying models for estimators like $\bar{y}$, $\hat{\bar{y}}_R$, $\hat{\bar{y}}_{lr}$ etc. are well known in survey sampling. Better understanding of the model for which each estimator is used would help samplers to choose the 'right' estimator. A finite population decomposition will be introduced to study the effect of the 'model deviation' on the performance of the estimators like mean-per-unit($\bar{y}$), ratio($\hat{\bar{y}}_R$) and regression($\hat{\bar{y}}_{lr}$) in Section 5. Using the same decomposition, we can rewrite the variance formulae of $\bar{y}$, $\hat{\bar{y}}_R$ and $\hat{\bar{y}}_{lr}$ in terms of some simple and interpretable population characteristics. The efficiency comparison can be easily made. In particular, we show that $\hat{\bar{y}}_{lr}$ is more efficient than $\hat{\bar{y}}_R$ and they are equally efficient if the intercept of the population regression line is zero. We also prove that under simple random sampling scheme the unweighted regression estimator is the most efficient estimator. A natural extension to p variates is given in Section 6. We show that the unweighted multivariate regression estimator is more efficient than the weighted ones. An intuitive argument is also given.

2. Bias of the Ratio Estimator

In general, the ratio estimator has bias and variance of the same order n^{-1}. Hence, in practice the bias usually is not so important in large samples. For small sample problems, such as in stratified sampling with many strata where the separate ratio estimator is used in each stratum with small sample size, the problem of bias of the ratio becomes somewhat important. Cochran(1977) pointed out that in surveys with many strata and small samples in each stratum if the separate ratio estimator seems appropriate, it may be useful to modify the ratio estimator such that it is unbiased or subject to a smaller order bias than the ratio.

Hartley and Ross(1954) proposed an unbiased estimator

$$\hat{\bar{y}}_{HR} = \bar{r}\,\bar{X} + \frac{n(N-1)}{n-1}(\bar{y} - \bar{r}\,\bar{x}), \tag{2.1}$$

where

$$\bar{r} = \frac{1}{n}\sum_1^n r_i = \frac{1}{n}\sum_1^n \frac{y_i}{x_i}. \tag{2.2}$$

Mickey(1959) extended Hartley and Ross 's ideas to get another unbiased estimator

$$\hat{\bar{y}}_M = \hat{\bar{R}}\,\bar{X} + \frac{n(N-n+1)}{N}(\bar{y} - \hat{\bar{R}}\,\bar{x}), \tag{2.3}$$

where

$$\hat{\bar{R}} = \frac{1}{n}\sum_1^n \frac{n\bar{y} - y_i}{n\bar{x} - x_i}. \tag{2.4}$$

Lahiri(1951) showed that the ordinary ratio estimator is unbiased under an unequal probability sampling scheme.

There are several methods available for reducing the bias to order n^{-2}. The first is the jackknife estimator of $\hat{\bar{y}}_R$, due to Quenouille(1956). Durbin(1959) is the

first one to propose the jackknife method for ratio. It can be applied to a broad class of statistical problems in which the original estimator has a bias of order n^{-1}. Beale(1962) proposed the following estimator for bias reduction.

$$\hat{\bar{y}}_B = \bar{X} \frac{\bar{y} + \frac{1-f}{n} \frac{s_{xy}}{\bar{x}}}{\bar{x} + \frac{1-f}{n} \frac{s_x^2}{\bar{x}}} = \hat{\bar{y}}_R \frac{1 + \frac{1-f}{n} c_{xy}}{1 + \frac{1-f}{n} c_{xx}}, \tag{2.5}$$

where

$$c_{xy} = \frac{s_{xy}}{\bar{x}\,\bar{y}}, \; c_{xx} = \frac{s_x^2}{\bar{x}^2}. \tag{2.6}$$

Tin(1965) proposed an estimator which is closely related to Beale's estimator

$$\hat{\bar{y}}_T = \hat{\bar{y}}_R \left[1 - \frac{1-f}{n} (c_{xx} - c_{xy})\right]. \tag{2.7}$$

Tin's correction, i.e. the second term of (2.7), is a sample analog of the bias (Cochran, 1977, p.161)

$$E(\hat{\bar{y}}_R - \bar{Y}) = \frac{1-f}{n} (C_{xx} - C_{xy}) \bar{Y} + O(n^{-2}), \tag{2.8}$$

where

$$C_{xy} = \frac{S_{xy}}{\bar{X}\,\bar{Y}}, \; C_{xx} = \frac{S_x^2}{\bar{X}^2}. \tag{2.9}$$

Cochran(1977) pointed out $\hat{\bar{y}}_B$ and $\hat{\bar{y}}_T$ have the same leading term of order n^{-1}. In general , $\hat{\bar{y}}_B$ and $\hat{\bar{y}}_T$ should perform very similarly for large sample.

We will consider a decomposition of the finite population. Using this decomposition, we can give an interpretation of the bias in (2.8). In fact, we will see that the leading term of the bias of $\hat{\bar{y}}_R$ is caused by the non-zero intercept of the regression line to the finite population.

Given a finite population $\{(y_i, x_i), i=1,..N\}$, we can decompose the population as following

$$y_i = \alpha + \beta x_i + e_i , \tag{2.10}$$

where

$$\beta = \frac{\sum_1^N (x_i - \bar{X})(y_i - \bar{Y})}{\sum_1^N (x_i - \bar{X})^2} , \tag{2.11}$$

$$\alpha = \bar{X} (R - \beta) . \tag{2.12}$$

It is easy to see $\{ e_i\}$ satisfies

$$\sum_1^N e_i = 0 , \ \sum_1^N e_i x_i = 0 . \tag{2.13}$$

Using (2.10), we can find the leading term of the bias of $\hat{\bar{y}}_R$ in terms of the intercept, α

$$E(\hat{\bar{y}}_R - \bar{Y}) = \frac{1-f}{n} \alpha C_{xx} + O(n^{-2}) , \tag{2.14}$$

which follows from (2.10), (2.13), and $C_{xy} = \frac{S_{xy}}{\bar{X} \bar{Y}} = \frac{\beta}{R} C_{xx}$.

Formula (2.14) shows that the bias of $\hat{\bar{y}}_R$ is caused by the non-zero intercept α in the decomposition. Furthermore, the leading term of the bias depends on αC_{xx}. Since $C_{xx} > 0$, the sign of the bias is the same as the sign of α. That is, if $\alpha > 0$ then we would expect that $\hat{\bar{y}}_R$ will slightly overestimate $\bar{Y}$; if $\alpha < 0$ then $\hat{\bar{y}}_R$ will underestimate $\bar{Y}$. There is no certainty that for small n the actual bias is reduced. However, we have reason to expect that the bias will be diminished when n is not too small and the population is not extremely irregular.

3. Bias of the Regression Estimator

Like the ratio estimator, the linear regression estimator, $\hat{\bar{y}}_{lr}$, is also a biased estimator. The leading term of the bias is given as follows (Cochran 1977, p.198)

$$E(\hat{\bar{y}}_{lr} - \bar{Y}) = - \frac{1-f}{n} \frac{E(e_i(x_i - \bar{X})^2)}{S_x^2} + O(n^{-2}), \tag{3.1}$$

where e_i is defined in (2.10)

To provide an interpretation for the leading term of the bias of $\hat{\bar{y}}_{lr}$, we use the following quadratic decomposition of the finite population

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + d_i, \tag{3.2}$$

where $\{ \beta_j , j=1,2,3\}$ minimizes

$$\sum_1^N (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2 .$$

That is

$$(\beta_0, \beta_1, \beta_2)' = (X'X)^{-1}X'y , \tag{3.3}$$

where

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_N & x_N^2 \end{bmatrix} , \quad y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{bmatrix} . \tag{3.4}$$

From the least squares theory, it is easy to see

$$\sum_1^N d_i = 0 , \sum_1^N d_i x_i = 0 , \sum_1^N d_i x_i^2 = 0 . \tag{3.5}$$

Using the decomposition (3.2), we can show

Theorem 3.1. Let $\beta_0, \beta_1, \beta_2$ as in (3.3) and $\bar{X}^{(t)} = \frac{1}{N}\sum_1^N x_i^t$, the t-th population moments of x, $\bar{x}^{(t)}$ be the t-th sample moments, then

$$E(\hat{\bar{y}}_{lr} - \bar{Y}) = -\frac{1-f}{n}\,\beta_2\,\frac{K}{S_x^2} + O(n^{-2}),$$

where

$$K = [\,\bar{X}^{(4)} - (\bar{X}^{(2)}, \bar{X}^{(3)}) \begin{pmatrix} 1 & \bar{X} \\ \bar{X} & \bar{X}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} \bar{X}^{(2)} \\ \bar{X}^{(3)} \end{pmatrix}] = \frac{\begin{vmatrix} 1 & \bar{X} & \bar{X}^{(2)} \\ \bar{X} & \bar{X}^{(2)} & \bar{X}^{(3)} \\ \bar{X}^{(2)} & \bar{X}^{(3)} & \bar{X}^{(4)} \end{vmatrix}}{\begin{vmatrix} 1 & \bar{X} \\ \bar{X} & \bar{X}^{(2)} \end{vmatrix}} > 0 .$$

Proof. We can rewrite (2.10) in the matrix form

$$y = X H \underset{\sim}{\alpha} + e, \tag{3.6}$$

where

$$H = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \underset{\sim}{\alpha} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

$e = (e_1, e_2, .., e_N)'$, X and y defined in (3.4). Using (2.13), (3.3) and (3.6),

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ 0 \end{pmatrix} + (\frac{1}{N} X'X)^{-1} \begin{pmatrix} \frac{1}{N}\sum_1^N e_i \\ \frac{1}{N}\sum_1^N e_i x_i \\ \frac{1}{N}\sum_1^N e_i x_i^2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ 0 \end{pmatrix} + \begin{pmatrix} * \\ * \\ c\,\frac{1}{N}\sum_1^N e_i x_i^2 \end{pmatrix}. \tag{3.7}$$

Therefore

$$\beta_2 = c\,E(e_i x_i^2), \tag{3.8}$$

where the explicit expression of c can be found using the formula for the matrix inversion

$$c = \frac{\begin{vmatrix} 1 & \bar{X} \\ \bar{X} & \bar{X}^{(2)} \end{vmatrix}}{\begin{vmatrix} 1 & \bar{X} & \bar{X}^{(2)} \\ \bar{X} & \bar{X}^{(2)} & \bar{X}^{(3)} \\ \bar{X}^{(2)} & \bar{X}^{(3)} & \bar{X}^{(4)} \end{vmatrix}} = [\, \bar{X}^{(4)} - (\, \bar{X}^{(2)}, \bar{X}^{(3)}) \begin{pmatrix} 1 & \bar{X} \\ \bar{X} & \bar{X}^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} \bar{X}^{(2)} \\ \bar{X}^{(3)} \end{pmatrix}]^{-1} = K^{-1} \,. \quad (3.9)$$

Note in writing the last equality, we used the formula for the determinant of a partitioned matrix (e.g. Rao, 1971, p. 32). Using (3.9) and the fact that $[\frac{1}{N}(X'X)]$ is positive definite , we have

$$K = c^{-1} > 0 \,. \quad (3.10)$$

From (2.13), we have

$$E(\, e_i\, x_i^2) = E(\, e_i(\, x_i - \bar{X})^2) \,. \quad (3.11)$$

Theorem 3.1 follows from (3.1),(3.8),(3.10) and (3.11). □

Theorem 3.1 may provide a better understanding of the bias of $\hat{\bar{y}}_{lr}$. The leading term of the bias is due to the non-zero β_2 the coefficient of the quadratic term in the decomposition (3.2). Furthermore, we can see the over or underbias of $\hat{\bar{y}}_{lr}$ depends only on the sign of β_2. If $\beta_2 > 0$, then we expect $\hat{\bar{y}}_{lr}$ will underestimate $\bar{Y}$, whereas $\beta_2 < 0$ indicates an overestimation of $\hat{\bar{y}}_{lr}$.

4. Bias of the Multiple Regression Estimator

The discussion so far has been restricted to the situation in which auxiliary information on just one x-variate is to be used for improving the precision of estimates. In practice, we may have information about several x-variates and it may be considered important to make use of all the available information to get a more precise estimator. Several methods of using p-variates $X_1, X_2, \cdots, X_p$ are proposed in

the literature. The most popular estimator is the multivariate regression estimator

$$\hat{\bar{y}}_{mlr} = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j (\bar{x}_j - \bar{X}_j) , \qquad (4.1)$$

where $\{ \hat{\beta}_j \; j=1,..,p\}$ is the least squares estimate of the corresponding population parameter $\{ \beta_j, j=1..,p \}$ in the linear regression model. $\hat{\bar{y}}_{mlr}$ is the best linear unbiased estimator under the following superpopulation model(Royall, 1970)

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \varepsilon_i , \qquad (4.2)$$

where

$$E_M (\varepsilon_i) = 0 \; ; \; E_M (\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 w_i & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$$

with $w_i = 1$. Clearly, the multivariate regression is a consistent estimator of $\bar{Y}$. Like the ratio estimator and simple regression estimator, $\hat{\bar{y}}_{mlr}$ has also a bias of order n^{-1} . For p=2, Konijn(1973) gave an explicit but complicated expression of the leading term of the bias for the bivariate regression estimator, i.e.

$$\frac{1-f}{n} \frac{1}{1 - \rho^2} \left[\frac{2 S_{e12} \rho}{S_{X_1} S_{X_2}} - \frac{S_{e11}}{S_{X_1}^2} - \frac{S_{e22}}{S_{X_2}^2} \right] , \qquad (4.3)$$

where

$$S_{e12} = \frac{1}{N-1} \sum_1^N e_i (x_{1i} - \bar{X}_1)(x_{2i} - \bar{X}_2) ,$$

$$e_i = (y_i - \bar{Y}) - B_1 (x_{1i} - \bar{X}_1) - B_2 (x_{2i} - \bar{X}_2) ,$$

B_1 and B_2 are the population regression coefficient of y over X_1, X_2, and ρ is the correlation coefficient between X_1 and X_2 , and S_{X_1}, S_{X_2} are the population variances of X_1, X_2 respectively.

Our purpose is to find a simple formula for the leading term of the bias of $\hat{\bar{y}}_{mlr}$ and $\hat{\bar{y}}_w$ which is in a more general class of weighted multivariate regression estimators

$$\hat{\bar{y}}_w = (\bar{X}_0, \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)(X_s' W_s^{-1} X_s)^{-1} X_s' W_s^{-1} y_s$$

$$= \frac{1}{N} 1_N' X (X_s' W_s^{-1} X_s)^{-1} X_s' W_s^{-1} y_s ,$$

where

$$X = \begin{bmatrix} x_{01} & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{0N} & x_{1N} & \cdots & x_{pN} \end{bmatrix}, \quad X_s = \begin{bmatrix} x_{01} & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{0n} & x_{1n} & \cdots & x_{pn} \end{bmatrix},$$

$$\bar{X}_j = \frac{1}{N} \sum_1^N x_{ji} , \quad y_s = (y_1, y_2, \ldots, y_n)' , \quad 1_N = (1,1, \cdots ,1)'$$

and

$$W_s = \mathrm{diag}(w_1, w_2, \ldots, w_n)$$

is a sub-matrix of

$$W = \mathrm{diag}(w_1, w_2, \ldots, w_N) .$$

From the Theorem 1 of Wright(1983), we know that $\hat{\bar{y}}_w$ is a consistent estimator of $\bar{Y}$ if

$$1_N \in \mathrm{col}(W^{-1}X) , \text{ i.e. } W = X \underset{\sim}{c} , \text{ for some } \underset{\sim}{c} . \tag{4.4}$$

We will consider the class of estimators $\hat{\bar{y}}_w$ satisfying this condition. If $\hat{\bar{y}}_w$ contains the intercept term, then $x_{0i} \equiv 1$, otherwise X_0 should be omitted from $\hat{\bar{y}}_w$. It is easy to see that $\hat{\bar{y}}_R$, $\hat{\bar{y}}_{lr}$ and $\hat{\bar{y}}_{mlr}$ are special cases of $\hat{\bar{y}}_w$.

Using the weighted least square notation, we have

$$\hat{\bar{y}}_w = (\bar{X}_0, \bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)\,\hat{\underset{\sim}{\beta}}_w \equiv \bar{\underset{\sim}{X}}\,\hat{\underset{\sim}{\beta}}_w ,$$

where

$$\hat{\underset{\sim}{\beta}}_w = (X_s' W_s^{-1} X_s)^{-1} X_s' W_s^{-1} y_s . \tag{4.5}$$

The following decomposition of the finite population will be used to prove our key result

$$y = X \underset{\sim}{\beta}_w + e , \tag{4.6}$$

where

$$\underset{\sim}{\beta}_w = (X'W^{-1}X)^{-1} X'W^{-1}y . \tag{4.7}$$

It is easy to see that

$$X' W^{-1}e = 0 . \tag{4.8}$$

Lemma 4.1. Let $\hat{\underset{\sim}{\beta}}_w$, $\underset{\sim}{\beta}_w$ be defined as in (4.5),(4.7), then

$$\hat{\underset{\sim}{\beta}}_w - \underset{\sim}{\beta}_w = (X_s' W_s^{-1} X_s)^{-1} X_s' W_s^{-1} e_s = O_p(n^{-0.5})$$

$$= S_{xx,w}^{-1}\,\bar{\underset{\sim}{u}} + O_p(n^{-1}) , \tag{4.9}$$

where

$$S_{xx,w} = \frac{1}{N} X' W^{-1}X , \quad \bar{\underset{\sim}{u}} = (\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_p)' \tag{4.10}$$

and

$$\bar{u}_j = \frac{1}{n} \sum_1^n u_{ji} , \quad u_{ji} = x_{ji} w_i^{-1} e_i . \tag{4.11}$$

Proof. Since

$$\hat{\underset{\sim}{\beta}}_w = \underset{\sim}{\beta}_w + (X_s' W_s^{-1} X_s)^{-1} X_s' W_s^{-1} e_s , \qquad (4.12)$$

it is easy to see

$$\frac{1}{n} X_s' W_s^{-1} X_s = \frac{1}{N} X'W^{-1}X + O_p(n^{-0.5}) \equiv S_{xx,w} + O_p(n^{-0.5}) . \qquad (4.13)$$

And from (4.8), we have

$$\frac{1}{n} X_s' W_s^{-1} e_s = (\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_p)' = O_p(n^{-0.5}) . \qquad (4.14)$$

Hence, using (4.13) and (4.14)

$$\hat{\underset{\sim}{\beta}}_w - \underset{\sim}{\beta}_w = S_{xx,w}^{-1} \underset{\sim}{\bar{u}} + O_p(n^{-1}) .$$

This completes the proof of Lemma 4.1. □

Theorem 4.1. If $1_N \in \text{col}(W^{-1}X)$, then

$$E(\hat{\bar{y}}_w - \bar{Y}) = - \frac{1-f}{n} \text{tr}(S_{xx,w}^{-1} S_{xu,w}) + O(n^{-2}) ,$$

where tr(A) denotes the trace of the matrix A and $S_{xx,w}$ is defined in (4.10),

$$S_{xu,w} = \left[S_{u_j x_k} \right]$$

and

$$S_{u_j x_k} = \frac{1}{N-1} \sum_1^N u_{ji}(x_{ki} - \bar{X}_k) = \frac{1}{N-1} \sum_1^N x_{ji} e_i w_i^{-1}(x_{ki} - \bar{X}_k) .$$

Proof. Let $\underset{\sim}{\bar{x}} = 1_n' X_s / n$ and $\underset{\sim}{\bar{X}} = 1_N'X / N$. Since $1_N \in \text{col}(W^{-1}X)$, we have

$\bar{y} = \underset{\sim}{\bar{x}} \hat{\underset{\sim}{\beta}}_w$, which together with Lemma 4.1 implies

$$\hat{\bar{y}}_w = [\underset{\sim}{\bar{x}} - (\underset{\sim}{\bar{x}} - \underset{\sim}{\bar{X}})] \hat{\underset{\sim}{\beta}}_w = \bar{y} - (\underset{\sim}{\bar{x}} - \underset{\sim}{\bar{X}})[\underset{\sim}{\beta}_w + S_{xx,w}^{-1} \underset{\sim}{\bar{u}} + O_p(n^{-1})]$$

$$= \bar{y} - (\underset{\sim}{\bar{x}} - \underset{\sim}{\bar{X}}) \underset{\sim}{\beta}_w - (\underset{\sim}{\bar{x}} - \underset{\sim}{\bar{X}}) S_{xx,w}^{-1} \underset{\sim}{\bar{u}} + O_p(n^{-1.5}) .$$

Hence, we have

$$E(\hat{\bar{y}}_w - \bar{Y}) = -E((\bar{\underline{x}} - \bar{\underline{X}})S_{xx,w}^{-1}\bar{\underline{u}}) + O(n^{-2}).$$

And

$$E((\bar{\underline{x}} - \bar{\underline{X}})S_{xx,w}^{-1}\bar{\underline{u}}) = E(tr[(\bar{\underline{x}} - \bar{\underline{X}})S_{xx,w}^{-1}\bar{\underline{u}}]) = E(tr[S_{xx,w}^{-1}\bar{\underline{u}}(\bar{\underline{x}} - \bar{\underline{X}})])$$

$$= tr(S_{xx,w}^{-1}E[\bar{\underline{u}}(\bar{\underline{x}} - \bar{\underline{X}})]) = \frac{1-f}{n} tr(S_{xx,w}^{-1}S_{xu,w}).$$

In writing the above equalities, we use the fact Tr(AB)=Tr(BA) and

$$E(\bar{\underline{u}}(\bar{\underline{x}} - \bar{\underline{X}})) = \left(E[\bar{u}_j(\bar{x}_k - \bar{X}_k)]\right) = \frac{1-f}{n}\left(S_{u_j,x_k}\right).$$

This completes the proof of Theorem 4.1. □

Three special cases of Theorem 4.1 are mentioned below.

(1) For the ratio estimator, $X = (x_1, x_2, .., x_N)'$, $W = diag(x_1, x_2, .., x_N)$, $\beta_w = \bar{Y}/\bar{X} = R$ and $e_i = y_i - \beta_w x_i = y_i - R x_i$. In terms of the decomposition (2.10),

$$S_{xu,w} = \frac{1}{N-1}\sum_1^N e_i(x_i - \bar{X}) = S_x^2(\beta - R) = -\alpha\frac{S_x^2}{\bar{X}}.$$

Therefore, the leading term of the bias of $\hat{\bar{y}}_R$ is

$$-\frac{1-f}{n} tr(S_{xx,w}^{-1}S_{xu,w}) = \frac{1-f}{n}\alpha\frac{S_x^2}{\bar{X}^2},$$

which is the same as (2.14).

(2) For the regression estimator, we have

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, W = I_{NxN}.$$

It is easy to see

$$S_{xx,w} = \begin{bmatrix} 1 & \bar{X} \\ \bar{X} & \bar{X}^{(2)} \end{bmatrix} , \; S_{xu,w} = \begin{bmatrix} 0 & 0 \\ 0 & E(e_i x_i^2) \end{bmatrix} + O(N^{-1}) .$$

Hence, the leading term of the bias of $\hat{\bar{y}}_{lr}$ is

$$- \frac{1-f}{n} \, tr \, (S_{xx,w}^{-1} S_{xu,w}) = - \frac{1-f}{n} \frac{E(e_i x_i^2)}{S_x^2} .$$

(3) It can be verified that Konijn's(1973) expression of the bias for the special case with p=2 is the same as our formula in Theorem 4.1. Our formula is much more general than Konijn's even for p=2. Note that for the multivariate regression estimator, W = I and 1_N in the design matrix X, and

$$\sum_1^N e_i (x_{ki} - \bar{X}_k) = 0 ,$$

which implies

$$\frac{1}{N-1} \sum_1^N x_{ji} \, e_i (x_{ki} - \bar{X}_k) = \frac{1}{N-1} \sum_1^N (x_{ji} - \bar{X}_j) \, e_i (x_{ki} - \bar{X}_k) .$$

This is the same as the S_{e12}'s in Konijn's formula in (4.3).

5. Efficiency Comparison of $\bar{y}$, $\hat{\bar{y}}_R$ and $\hat{\bar{y}}_{lr}$

The variance of $\bar{y}$ is well-known,

$$Var(\bar{y}) = \frac{1-f}{n} \, S_y^2 = \frac{1-f}{n} \frac{1}{N-1} \sum_1^N (y_i - \bar{Y})^2 . \qquad (5.1)$$

There are no closed forms for $Var(\hat{\bar{y}}_R)$ and $Var(\hat{\bar{y}}_{lr})$. Each can be approximated by their approximate variances (Cochran, 1977)

$$V_R = \frac{1-f}{n} \frac{1}{N-1} \sum_1^N (y_i - \frac{\bar{Y}}{\bar{X}} x_i)^2 \qquad (5.2)$$

and

$$V_{lr} = \frac{1-f}{n} \frac{1}{N-1} \sum_1^N [(y_i - \bar{Y}) - B(x_i - \bar{X})]^2 , \qquad (5.3)$$

where $f = n/N$ is the sampling fraction and $B = \beta$ is given in (2.11).

We would like to rewrite the expression of V_R, V_{lr} and $Var(\bar{y})$ in terms of the decomposition in (2.10). Using the decomposition (2.10), we have

Theorem 5.1. Let α,β be defined as in (2.10), then

(a) $\hat{\bar{y}}_R - \bar{Y} = (1 - (\frac{\bar{X}}{\bar{x}}))(-\alpha) + (\frac{\bar{X}}{\bar{x}})\bar{e}$, where $\bar{e} = \frac{1}{n}\sum_1^n e_i$

(b) $V_R = \frac{1-f}{n}(\alpha^2 \frac{S_x^2}{\bar{X}^2} + S_e^2)$, where $S_e^2 = \frac{1}{N-1}\sum_1^N e_i^2$.

Proof. From (2.10) and (2.13), we have

$$\hat{\bar{y}}_R = \alpha(\frac{\bar{X}}{\bar{x}}) + \beta\bar{X} + \frac{\bar{e}}{\bar{x}}\bar{X} \quad \text{and} \quad \bar{Y} = \alpha + \beta\bar{X},$$

which implies Part(a). From (5.2), we have

$$V_R = \frac{1-f}{n} \frac{1}{N-1} \sum_1^N d_i^2 , \; d_i = -\alpha \frac{(x_i - \bar{X})}{\bar{X}} + e_i . \qquad (5.4)$$

Using (2.13), it is easy to see

$$\sum_1^N e_i(x_i - \bar{X}) = 0 . \qquad (5.5)$$

Part(b) follows immediately from (5.4)-(5.5). □

Part(a) of Theorem 5.1 shows that $\hat{\bar{y}}_R - \bar{Y}$ depends only on the intercept, α, not on the slope β. $\hat{\bar{y}}_R - \bar{Y}$ can be written as a "weighted" average of $-\alpha$ and $\bar{e}$, the weight for $\bar{e}$ is $\bar{X}/\bar{x}$. (Note that the weight $\bar{X}/\bar{x}$ may be greater than 1). For a sample with $\bar{x} = \bar{X}$, the weight associated with $\bar{e}$ is 1, and zero weight for $-\alpha$. Hence, there is no effect of α on the difference $\hat{\bar{y}}_R - \bar{Y}$ for any sample

with $\bar{x} = \bar{X}$. Such a sample is called "balanced sample" in Royall and Herson(1973).

From Part(b) of Theorem 5.1, we can see that V_R, the leading term of $\mathrm{Var}(\hat{\bar{y}}_R)$, is composed of two sources of variation. The first component is the non-zero intercept (α) of the decomposition and the population characteristic ($S_x^2/\bar{X}^2$) of x-variate. The second is the population variance (S_e^2) of e-variate. For $\bar{y}$, the mean-per-unit estimator, we have

Theorem 5.2.

(a) $\bar{y} - \bar{Y} = \beta(\bar{x} - \bar{X}) + \bar{e}$.

(b) $\mathrm{Var}(\bar{y}) = \frac{1-f}{n}(\beta^2 S_x^2 + S_e^2)$.

(c) $V_R \le \mathrm{Var}(\bar{y})$ if and only if $|\alpha| \le |\beta| \bar{X}$,
where α and β are defined in (2.11),(2.12).

Proof. Part(a) is trivial. Using (2.10) and (2.13), we have

$$y_i - \bar{Y} = \beta(x_i - \bar{X}) + e_i . \qquad (5.6)$$

Part(b) follows from (5.1), (5.5) and (5.6). Part(c) follows from Part(b) and Theorem 5.1. □

For the regression estimator, $\hat{\bar{y}}_{lr}$, we have

Theorem 5.3.

(a) $\hat{\bar{y}}_{lr} - \bar{Y} = \bar{e} + O_p(n^{-1})$

(b)$V_{lr} = \frac{1-f}{n} S_e^2$

(c)$V_{lr} \leq V_R$, with strict inequality if $\alpha \neq 0$

(d)$V_{lr} \leq \mathrm{Var}(\bar{y})$, with strict inequality if $\beta \neq 0$

Proof. Parts(a) and (b) are trivial. Part(c) follows from Theorem 5.1. Part(d) follows from Theorem 5.2. □

From Part(b) of Theorem 5.3, we see that V_{lr} does not depend on α and β, whereas V_R depends on the intercept α and $\mathrm{Var}(\bar{y})$ depends on the slope β. The reason is that the underlying model for $\hat{\bar{y}}_{lr}$ has the same structure as the decomposition (2.10). On the other hand, the underlying model for $\hat{\bar{y}}_R$ is

$$y_i = \beta x_i + \varepsilon_i ,$$

hence $\hat{\bar{y}}_R$ captures the slope term in the decomposition but not the intercept term. The underlying model for $\bar{y}$ is

$$y_i = \alpha + \varepsilon_i ,$$

hence the intercept in (2.10) can be captured by $\bar{y}$ but not the slope.

Part(c) of Theorem 5.3 shows that $\hat{\bar{y}}_{lr}$ is always more efficient than $\hat{\bar{y}}_R$. Note the result of Part(c) is well-known(Cochran, 1977, p.196) without using the finite population decomposition approach. For estimating cell totals in tables of the type typically constructed from survey data, Fuller(1977) showed the superior performance of the regression estimator. However, in practice $\hat{\bar{y}}_R$ is more popular than $\hat{\bar{y}}_{lr}$. One reason is the computational simplicity of $\hat{\bar{y}}_R$ over $\hat{\bar{y}}_{lr}$ in complex situations.

Wright(1983) characterized a class of consistent estimators for a general sampling plan. Applying his result to the simple random sampling, we can see

$$\hat{\bar{y}}_w = \hat{\alpha}_w + \hat{\beta}_w \bar{X} \tag{5.7}$$

is a consistent estimator of $\bar{Y}$ if w_i is chosen to be either 1's , x_i's or $c_1 + c_2 x_i$, where

$$(\hat{\alpha}_w , \hat{\beta}_w)' = (X_s' W_s^{-1} X_s)^{-1} X_s' W_s^{-1} y_s ,$$

$$X_s = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad y_s = (y_1, y_2, .., y_n)' , \quad W_s = \mathrm{diag}(w_1, w_2, \cdots, w_n) .$$

However, it is not clear that how to choose the "optimal" weight. One criterion we may use is the minimum variance of $\hat{\bar{y}}_w$. According to which, we can choose the "best" weight among this class of consistent estimators of $\bar{Y}$. Note that $\hat{\bar{y}}_{lr}$ is also a special case of $\hat{\bar{y}}_w$ with $w_i = 1$. Theorem 5.4 shows that $\hat{\bar{y}}_{lr}$ ($w_i = 1$) is the best choice.

Theorem 5.4. Let V_W denote the leading term of $\mathrm{Var}(\hat{\bar{y}}_w)$ and α, β be defined as in (2.10). If $w_i = c_1 + c_2 x_i$ for some c_1, c_2, then

(a) $V_W = \dfrac{1-f}{n} \dfrac{1}{N-1} \sum_1^N (y_i - \alpha_w - \beta_w x_i)^2$.

(b) For any α_0, β_0,

$$\sum_1^N (y_i - \alpha_0 - \beta_0 x_i)^2 = \sum_1^N (y_i - \alpha - \beta x_i)^2 + \sum_1^N [(\alpha_0 - \alpha) + (\beta_0 - \beta x_i)]^2$$

(c) $V_{lr} \le V_W$ for any choice of w_i . where $(\alpha_w, \beta_w)' = (X'W^{-1}X)^{-1} X'W^{-1}y$,

$$X = \begin{bmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_N \end{bmatrix} , \quad y = (y_1, y_2, \ldots, y_N)' , \quad W = \mathrm{diag}(w_1, w_2, \cdots, w_N) .$$

Proof. Since Theorem 5.4 will be a special case of Theorem 6.1 in Section 6, the proof is omitted.

From Part(c), we see that $\hat{\bar{y}}_{lr}$ (with $w_i = 1$) is the most efficient estimator among the class of consistent estimators $\hat{\bar{y}}_w$. This provides a strong justification for the use of $\hat{\bar{y}}_{lr}$ in survey sampling. Obviously, we would expect Theorem 5.4 also holds for multiple regression estimator when there are more than one auxiliary variables. We will extend Theorem 5.4 to the multivariate case in Section 6.

6. Optimal Weighted Multiple Regression Estimator

In this section we consider the efficiency of the estimators based on the p-auxiliary variables. Consider the multivariate regression estimator, $\hat{\bar{y}}_{mlr}$, and $\hat{\bar{y}}_w$, which are defined Section 4.

There is no exact formula for Var($\hat{\bar{y}}_w$) and MSE($\hat{\bar{y}}_w$). It can be easily shown that the leading terms of Var($\hat{\bar{y}}_w$) and MSE($\hat{\bar{y}}_w$) are the same. (Note that it is true only for the case $1_N \in \mathrm{col}(W^{-1}X)$). Let V_W denote the leading term of Var($\hat{\bar{y}}_w$). Lemma 6.1 finds an expression for V_W .

Lemma 6.1. If $1_N \in \mathrm{col}(W^{-1}X)$, then

$$V_W = \frac{1-f}{n} \frac{1}{N-1} \sum_1^N e_i^2 , \text{ where } e_i = y_i - \underset{\sim}{X}_i \underset{\sim}{\beta}_w \tag{6.1}$$

and

$$\underset{\sim}{\beta}_w = (X'W^{-1}X)^{-1} X'W^{-1}y . \tag{6.2}$$

Proof. Let $\underset{\sim}{\overline{X}} = 1_N'X / N$ and $\underset{\sim}{\overline{x}} = 1_n' X_s / n$. Since $1_N \in \text{col}(W^{-1}X)$, we have

$$\hat{\overline{y}}_w - \overline{Y} = (\overline{y} - \overline{Y}) - (\underset{\sim}{\overline{x}} - \underset{\sim}{\overline{X}}) \underset{\sim}{\beta}_w + O_p(n^{-1}) . \tag{6.3}$$

From (4.4), it is easy to see

$$\underset{\sim}{\overline{X}} \underset{\sim}{\beta}_w = \underset{\sim}{\overline{X}}(X'W^{-1}X)^{-1} X'W^{-1}y = \overline{Y} . \tag{6.4}$$

Hence

$$\hat{\overline{y}}_w - \overline{Y} = \overline{y} - \underset{\sim}{\overline{x}} \underset{\sim}{\beta}_w + O_p(n^{-1}) = \overline{e} + O_p(n^{-1}) , \tag{6.5}$$

where

$$\overline{e} = \frac{1}{n} \sum_1^n e_i , \quad e_i = y_i - \underset{\sim}{X}_i \underset{\sim}{\beta}_w . \tag{6.6}$$

Note that

$$\frac{1}{N} \sum_1^N e_i = \frac{1}{N} 1_N'(y - X \underset{\sim}{\beta}_w) = \overline{Y} - \underset{\sim}{\overline{X}} \underset{\sim}{\beta}_w = 0 .$$

Therefore

$$\overline{e} = O_p(n^{-0.5}) \tag{6.7}$$

and

$$E[(\hat{\overline{y}}_w - \overline{Y})^2] = \frac{1-f}{n} S_e^2 + O(n^{-2}) .$$

This completes the proof of Lemma 6.1. □

Using Lemma 6.1, we can easily prove the key result

Theorem 6.1. Let V_{lr}, V_W denote the leading terms of $Var(\hat{\bar{y}}_{mlr}), Var(\hat{\bar{y}}_w)$ and $\underset{\sim}{\beta} = (X'X)^{-1}X'y$ For any $\underset{\sim}{\beta}_0$, we have

(a)$(y - X\underset{\sim}{\beta}_0)'(y - X\underset{\sim}{\beta}_0) = (y - X\underset{\sim}{\beta})'(y - X\underset{\sim}{\beta}) + (\underset{\sim}{\beta}_0 - \underset{\sim}{\beta})'X'X(\underset{\sim}{\beta}_0 - \underset{\sim}{\beta})$.

(b)$V_{lr} \le V_W$. i.e. $\hat{\bar{y}}_{mlr}$ (with $w_i = 1$) is the most efficient estimator among $\hat{\bar{y}}_w$.

Proof. Define

$$e = y - X\underset{\sim}{\beta}_0 = (y - X\underset{\sim}{\beta}) + X(\underset{\sim}{\beta} - \underset{\sim}{\beta}_0) = d + X(\underset{\sim}{\beta} - \underset{\sim}{\beta}_0). \tag{6.8}$$

Since $d'X = (y - X\underset{\sim}{\beta})'X = y'(I - X(X'X)^{-1}X')X = 0$, we have

$$e'e = d'd + (\underset{\sim}{\beta} - \underset{\sim}{\beta}_0)'X'X(\underset{\sim}{\beta} - \underset{\sim}{\beta}_0). \tag{6.9}$$

This is Part(a). Part(b) follows easily from Part(a) (with $\underset{\sim}{\beta}_0 = \underset{\sim}{\beta}_w$), Lemma 6.1, $V_{lr} = \frac{1-f}{n}\frac{1}{N-1}\sum_1^N d_i^2$ and $V_W = \frac{1-f}{n}\frac{1}{N-1}\sum_1^N e_i^2$. □

A more intuitive proof of Theorem 6.1 is given below: From Lemma 6.1, we have

$$Var(\hat{\bar{y}}_w) = V_W = \frac{1-f}{n}\frac{1}{N-1}\sum_1^N (y_i - \underset{\sim}{X}_i\underset{\sim}{\beta}_w)^2, \tag{6.10}$$

which is minimized by taking $\underset{\sim}{\beta}_w$ to be the **unweighted** least squares estimate since (6.10) is an **unweighted** sum of squares. For unequal sampling scheme, (6.10) might be a weighted least squares. In that case $w_i = 1$ is no longer the optimal choice. We have shown that the most efficient estimator can be obtained by choosing W=I. However, it is interesting to see the comparison of $\hat{\bar{y}}_w$ under different criteria, for example, the coverage probabilities of the associated t-intervals and so on.

REFERENCES

Beale, E. M. L. (1962),"Some uses of computers in operational research," *Industrielle Organisation*, 31, 51-52.

Cochran, W. G. (1977), *Sampling Techniques, 3rd edition.* New York: Wiley.

Durbin, J. (1959),"A note on the application of Quenouille's method of bias reduction to the estimation of ratios," *Biometrika, 46*, 477-480.

Fuller, W. A. (1977),"A note on regression estimation for sample surveys," unpublished manuscript.

Hartley, H. O. and Ross, A. (1954),"Unbiased ratio estimates," *Nature, 174*, 270-271.

Konijn, H. S. (1973), *Statistical Theory of Sample Survey Design and Analysis.* North-Holland Pub. Co.

Lahiri, D. B. (1951),"A method for sample selection providing unbiased ratio estimates," *Bull. Int. Stat. Inst., 33*, 133-140.

Mickey, M. R. (1959),"Some finite population unbiased ratio and regression estimators," *Journal of the American Statistical Association, 54*, 594- 612.

Quenouille, M. H. (1956),"Notes on bias in estimation," *Biometrika, 43*, 353-360.

Rao, C. R. (1971), *Linear Statistical Inference and its Applications, 2nd edition.* New York: John Wiley & Sons.

Royall, R. M. (1970),"On finite population sampling theory under certain linear regression models," *Biometrika, 57*, 377-387.

Royall, R. M. and Herson, J. (1973),"Robust estimation in finite populations, I," *Journal of the American Statistical Association, 68*, 880-889.

Tin, M. (1965),"Comparison of some ratio estimators," *Journal of the American Statistical Association, 60*, 294-307.

Wright, R. L. (1983),"Finite population sampling with multivariate auxiliary information," *Journal of the American Statistical Association, 78*, 879-884.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2870	2. GOVT ACCESSION NO. AD-A163	3. RECIPIENT'S CATALOG NUMBER 621
4. TITLE (and Subtitle) BIAS AND EFFICIENCY OF THE CONSISTENT WEIGHTED REGRESSION ESTIMATORS IN FINITE POPULATION SAMPLING		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Lih-Yuan Deng		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE September 1985 13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Ratio estimator; Regression estimator; Weighted regression estimator; Bias; Efficiency; Finite population decomposition.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The leading terms of the bias of the ratio and regression estimators are known to be of order n^{-1}. We use a finite population decomposition to give a different expression for the leading term of the bias. Fitting a regression line to the finite population, we show that the intercept of the regression line causes the bias of the ratio estimator. Fitting a quadratic regression to the finite population, we show that the bias of the regression estimator is caused by the quadratic term. We also give a compact and intuitive formula for the leading term of the bias of the weighted regression estimators for

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT - cont'd.

p-auxiliary variables. Using the same decomposition, we can rewrite the variance formula of some popular estimators in terms of some simple and interpretable population characteristics. We prove that under simple random sampling scheme the unweighted regression estimator is the most efficient estimator. The extension for the p-auxiliary variates is also given.

END

FILMED

3 -86

DTIC